Indian Handwritten Script Identification System Based on Random Forest Tree Ensembles

Lalit P. Ganorkar, Dinesh V. Rojatkar

Abstract: ***The work proposal addresses to introduce a methodology for Indian unconstrained handwritten script identification by practicing distinct features and classifiers. By utilizing classifiers like RF, SVM, k-NN, and LDA for Indian script identification using statistical, geometric, and structural features. To preserve all the information present on handwritten documents such as historical, medieval, inscription, financial administration, public records, government archives, letters, land councils, various agreements, etc. in digitalize form needs textual document processing system (e.g. OCR). To build a precise and productive multi-script/language textual document processing system must have script identification. For this study use, total 1288 (line wise) samples of ten scripts use in India are collected from different persons of different gender, age, education and region (rural or urban). After successful training and testing, 81.8% and 0.252 accuracies and the OOB error rate are achieved by Random Forest respectively. And 77.8%, 73.5%, and 65.5% accuracy is achieved in SVM, k-NN and LDA classifiers respectively.***

Index Terms: ***Handwritten script identification, SVM, k-NN, LDA, Random Forest Tree Ensembles (RF).***

I. INTRODUCTION

In the field of pattern recognition and machine analysis, handwritten document analysis plays an important role. A handwritten document may contain one or more scripts. While documents analysis the script on the document may be changed from documents to documents. So to develop accurate and efficient multi-script document processing must require the script identification.

India is a country of 1.3 billion people having different culture and different languages. There is variation in language and communication skills in every 100 Km. So there are different scripts/language in common and official use. Out of these, 13 are official scripts and 23 are official languages are used by 29 states of India. Devanagari is the most used script in India. Some scripts are similar (e.g. Malayalam and Tamil/ Kannada and Telugu) in the scene of writing and communication. Feature-wise i.e. Shirorekha/ Matra/ Headline is common in scripts like Devanagari, Bangla and Gurumukhi [1]. For official purpose Latin script (i.e. English) is commonly preferred. Script identification can be done by segmenting the document by page-wise, paragraph wise, line-wise, word-wise or character-wise. This depends on script nature and structure of writing.

To retrieve information from any documents e.g. ancient, medieval, epigraphs, palm leaf manuscript, and even unregulated document formats written in multi-scripts/ languages requires an optical character recognition system (OCR). This multi-scripts OCR system is strongly script dependent [2]. Government of India started to digitize all the documents of government offices. Most of these documents are handwritten documents in various languages/scripts such as 7/12 extract, financial administration, public records, government archives, letters, land councils, various agreements, etc. So the government required a system that can mechanically detect the script/language and preserve all information present on these documents. To create such real-world application requires very high accuracy for script identification system. There are various problem and challenges in handwritten scripts such as the similar structure of scripts and unconstrained handwriting.

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author

Lalit P. Ganorkar*, Department of Electronics Engineering, Government College of Engineering Amravati, Amravati, India.
Dinesh V. Rojatkar, Department of Electronics Engineering, Government College of Engineering Amravati, Amravati, India.

II. RELATED WORK

Judith Hochberg proposed a system, where uses mean, standard deviation, skew of 5 connected component features and LDA classifier for six script/language identification. Researcher specified that document without fragmented ruling lines and characters give the greater classification accuracy [3]. In 2003 V. Singhal notify that the unconstrained nature of writing style, size of word and characters, the interword and interline spacing will introduce an error and unreliability in the system. To eliminate researcher introduce preprocessing and filtering techniques in [4].

In 2004, for postal automation in India, K. Roy proposed a scheme Indian handwritten script identification using word-wise segmentation in [5]. Water reservoir based concept feature is practical for script identification when words or characters are touching. Reservoir attribute is applicable to recognize the region of a word where most of the characters present i.e. busy zone. Due to the existence of the low small component feature, small word or low grade of documents cause miss recognition or rebuff by tree classifier. Anoop M. Namboodiri applies shirorekha strength and confidence features to identify Devanagari script [6]. G. G. Rajput and Anita H. B. extracted the feature by transforming an image from time to frequency (i.e. disparity in luminous or color across the image) domain in [7]. Information of image in the time domain is not prominent as compare to frequency domain information.

The researcher uses DCT and DWT coefficients as a feature and k-NN classifier determine the Euclidean distances between the test feature vector and stored features for eight script identification. In 2010, Mallikarjun Hangarge is used morphological filter to extract 13 spatial features (such as stroke density, pixel density, etc.). Also, pixels of less than 40 are eliminated from image to compute features [8]. Sk Md Obaidullah presents a set of 41 features that identify the six Indian scripts with the help of MLP classifier. Circularity is the most important feature among the mathematical feature and the fractal-based feature is the predominant feature among the structure based feature in [9].

In 2014, Rajmohan Pardeshi uses the feature based on spatial information and multi-resolution in [10], DCT coefficients of first 10 are preferred and added together to generate total 20 of features for script identification. Sub-band coding of DWT, projection of the RT, 46 dimensions of feature vector yields from SFs are used to compute standard deviation and entropy. Classification issue (i.e. 2 class problem) is solved using entropy and standard deviation in printed and handwritten text word.

In 2017, G. G. Rajput and Suryakant Baburao Ummapure propose a system at word level for script identification using scale-invariant feature transformation for feature extraction. STFT is used to avoid rotation and scale effect and to overcome from row shift effect; row shift invariant packet remodel is used [11].

In [12] S. Chanda and U. Pal use binary tree classifier algorithm and set of features i.e. a top reservoir, bottom reservoir, water flow level, reservoir baseline and height of reservoirs, head-line feature, vertical stroke distribution feature, and overlapping of a component feature, etc. for script identification. This proposed method is independent of text size and due to robust features is not rely on style and font of character in a word.

III. EXPERIMENTATION

This work proposed address to utilize RF, SVM, k-NN, and LDA classifiers for unconstrained handwritten ten Indian script identification by using statistical, geometric, and structural feature extraction. In these experiments, we are using the confusion matrix, ROC, and OOB error in RF as our evaluation metric to measure the execution of the technique.

A. Data Generation

To create a dataset for the present work 1288 samples are collected from different people of different gender, age, education, profession, and region (rural or urban). Variation in volunteers is seen in their handwriting style, font, strokes, and spacing. This will introduce complexity in the script identification for the system. As human being is also required deep observation to identify the similar scripts written in different handwriting. This database includes 10 scripts use in India for official or general purposes i.e. Bangla, Devanagari, English, Gujarati, Gurumukhi, Kannada, Malayalam, Tamil, Telugu, and Urdu. Certain specifications are kept constant for data generation such as color format, file format, and the resolution (i.e. 300 dpi). These specifications are confirmed after evaluating the effect of it on the system.

B. Image Pre-processing

The focus of image pre-processing is to create dataset compatible with a script identification system. We principally work together with general operation i.e. morphological operation, image binarization, and normalization. In this process of image pre-processing at first, we examine the image plans (i.e. RGB-3D numeric array) and found that an unnecessary image plane affects the system performance. So to enhance the efficiency of the system, image planes are converted to single plane i.e. grayscale by keeping luminance information and removing the saturation and hue. The grayscale value (GV) is calculated using the equation given below [13]

$$GV = 0.2989 * R + 0.5870 * G + 0.1140 * B \qquad (1)$$

Where, GV - Grayscale value,
R - Red component present in the image,
G - Green component present in the image, and
B - Blue component present in the image.

The binarization of an image is used to differentiate handwritten text or ink (foreground) and the background of the paper. For that Otsu's method [14] is selected to determine the threshold value. This method computes the threshold (e.g. 0.7176 thresholds is computed for fig.1), which is used to reduce the interclass variance of the white and black pixel. The function [15] we preferred to select threshold is passed over any non-zero imaginary part of the image. Due to handwriting style and use of ink may introduce certain unnecessary dot, lines, spots i.e. noise such as salt and pepper noise, marginal noise, and background noise, etc. in documents are removed during processing of an image. To enhance image quality binary image is inverted shows in fig. 1.

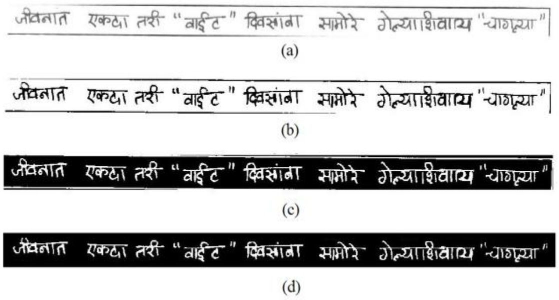


Fig. 1 Image Pre-processing. (a) Original Image. (b) Binary image. (c) Invert binary image. (d) Segmented image.

C. Feature Extraction

Features are significant facts about the script images, which are utilized to train a classifier. It is one of the most important factors that are preferred to enhance the execution system. In the case of handwriting, the same writer may write in a different pattern (i.e. shape and size) of characters depends on his mood, writing instruments (e.g. pen or pencil). This will create the difficulty for the system. The main purpose of feature descriptor is to lower down difficulty level and make easier for classifiers. The dimensionality of data affects the performance of the system. Smaller dimensionalities of the features are useful, especially for training a classifier in case of a small dataset. Several features utilized for handwritten script identification are discussed below.

1. Standard Deviation (S)

To measure dissimilarity or to evaluate the amount of variability or data set dispersion in statistic and image processing, the standard deviation is used. Standard deviation (S) is described mathematically [16],

$$S = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N} |I_i - \mu|^2} \qquad (2)$$

2. Entropy (E)

The texture of an image is characterized by evaluating entropy which is a measure of irregularity in an input image [17].

3. Kurtosis (k)

Kurtosis is computed the outline prone distribution or probability distribution of shape. This is fourth order moment and in image processing kurtosis value is a resolution and noise measurement. Here we use a measurement of resolution for script recognition and its classification [16].

4. Eccentricity

Eccentricity is preferred to measure the conical section differs from an actually circular shape. Two similar conical sections have equal eccentricity value.

5. Circularity

Circularity is preferred to limit the deviation of circular elements of a part surface. This feature is to distinguish the scripts which are written in a circular manner (i.e. Telugu, Urdu, etc.) and other scripts (i.e. English, Devanagari, etc.)

$$Circularity = \frac{Test\ shape\ area}{Area\ of\ the\ circle} \qquad (3)$$

6. Rectangularity

Standard measure to evaluate rectangularity of the script is as follows

$$Rectangularity = \frac{Ai}{Mn \times Mx} \qquad (4)$$

where,

Ai - mean of calculated area of image,
Mn - mean of minor axis length,
Mx - mean of major axis length.

7. Horizontal projection profile (HP)

This feature is useful for line wise script identification. Here in this study, we evaluate the standard deviation and max value of the horizontal projection.

8. Vertical projection profile (VP)

This feature is primarily useful for wordwise segmentation of script for recognition. But here in this study compute the standard deviation and max value of vertical projection. And we found these features are predominant over other features.

9. Aspect ratio

The aspect ratio of an image is calculated by dividing the width of the image with a height of the same image.

10. Euler Number

Euler number of a binary image is computed as the difference between the total number of objects and the total number of holes presents in those objects. Other Features such as area, perimeter, DCT and correlation coefficients are also used for script identification.

D. Classification

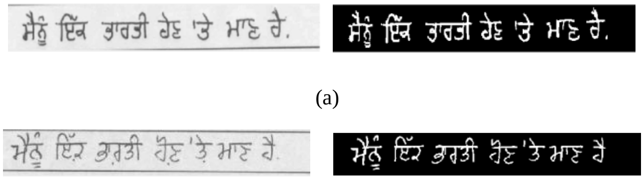
For classification purpose various machine learning algorithm like, SVM, k-NN, LDA and Random Forest (RF) are used. The details of each classifier are described as below:

SVM Classifier- Kernel function: Quadratic, Kernel scale: Automatic, Box constraint level: 1, Multiclass method: One-vs-One, Standardize data: true.

k-NN Classifier- Number of neighbors: 10, Distance metric: Euclidean, Distance weight: Squared inverse, Standardize data: true.

LDA Classifier- Discriminant type: Linear, Regularization: covariance.

RF Classifier- Number of trees: 300 'Number of predictors to sample: Square root of a number of variables.



(a)

(b)

Fig. 2 (a) Input and pre-process Gurumukhi script 1.
(b) Input and pre-process Gurumukhi script 2.

Gurumukhi script 1 and 2 from figure 2 is taken as input to the four classifiers i.e. RF, SVM, k-NN, and LDA. Script 2 is correctly identified as Gurumukhi script in all classifier, but script 1 is misclassified in RF and SVM classifier. In the case of k-NN and LDA classifier is correctly identified as Gurumukhi script.

In SVM classifier, script 1 in figure 2 is misclassified as Devanagari script; features which have more impact on this decision are eccentricity, perimeter, circularity, rectangularity, Euler number, aspect ratio, correlation coefficient, gray level and mean of image. Out of these Euler number plays an important role for that misclassification of Gurumukhi script as Devanagari script. This is due to strokes above headline (shirorekha) are more similar to the Devanagari script.

In RF classifier script 1 which is originally Gurumukhi script, misclassified as Tamil script. The reason behind this is that the writing style, vertical strokes below that headline are similar to Tamil script. Features like the mean of eccentricity and amount of variability in the perimeter are responsible for these misclassifications i.e. these features are very important for classification with respect to random forest classifier. Whereas LDA and k-NN are not taking too much importance to those features, so they classified correctly.

IV. EXPERIMENTAL RESULTS

Fig. 3 shows the evolution of individual features for RF and the vertical profile, covariance, and horizontal profile achieved the highest accuracy and lowest OOB error while Euler number has the lowest accuracy but highest OOB error. This graph shows the relation (app. inversely proportional) between OOB and accuracy. Fig. 4 shows the results obtained after integration of various features which are achieved outrageous accuracy and moderate OOB error. H, V, and C in the figure are horizontal, vertical projection profile and covariance respectively. H, C, and V are obtained good results independently. After integrating that features evaluation time is increased but with respective accuracy is not enhanced this clearly seen in this graph.

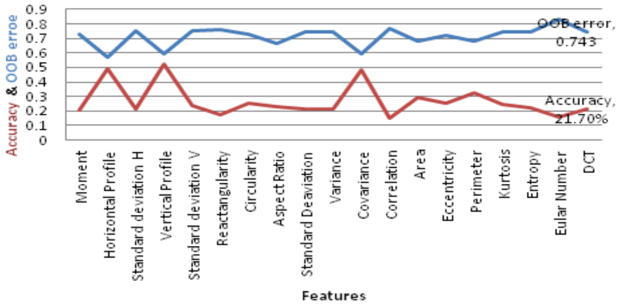
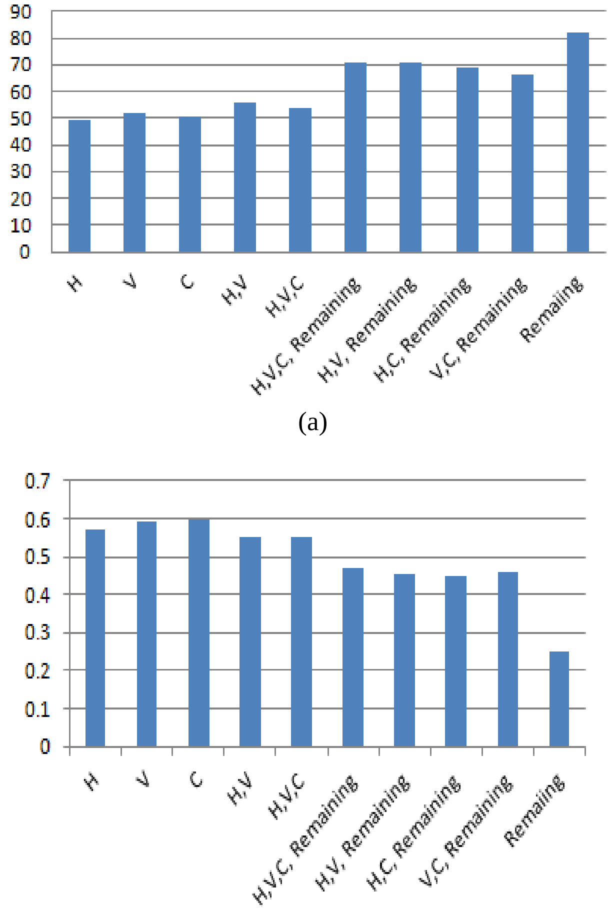


Fig. 3 Relation of features, accuracy and OOB error.



(a)

(b)

Fig. 4 Shows results after features integration. (a) accuracy. (b) OOB error rate.

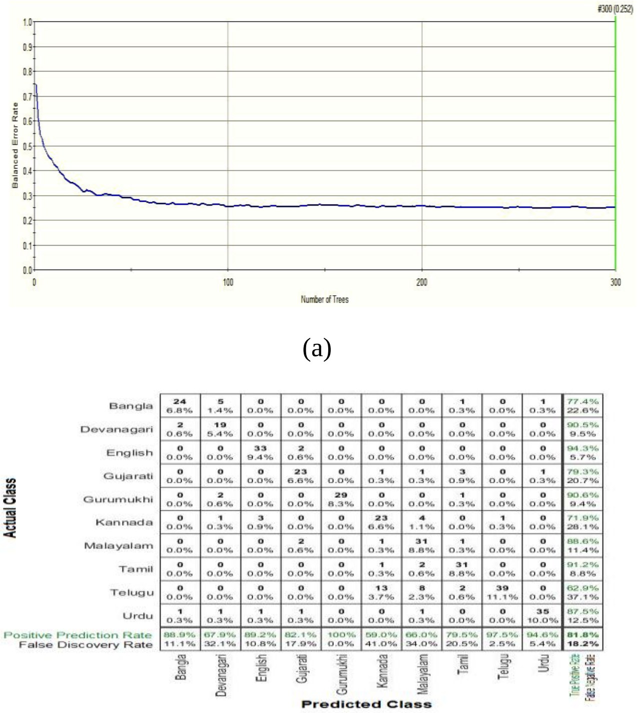
Table 1. shows some of the experimental results obtained for Random Forest. The results obtained at various levels during experimentation, some of them are mentioned here. OOB error is also known as an out-of-bag estimation. This is a method utilized for measuring the prediction error of random forests tree ensemble. Out-of-bag estimates aid avoid the need for an independent validation database. But often underrate the actual performance enhancement and the finest number of iterations [17]. Fig. 5 shows the OOB error rate of RF classifier, which is 0.252 for 300 trees and a confusion matrix having 81.8% accuracy for Indian scripts. Random Forest can handle high dimension feature vector and data set, reduce the variance and bias error and would not overfit easily.

Table 1. Experimental Results Obtained for RF of Script Identification

Features	Training/ Testing Samples	OOB error	Accuracy
Horizontal Profile, Vertical Profile, Area, Eccentricity, Perimeter, Rectangularity	700/ 88	0.54	54%
Horizontal Profile, Vertical Profile, Area, Eccentricity, Perimeter, Rectangularity, Standard Deviation, Euler Number, Entropy.	1000/ 288	0.49	57%
Horizontal Profile, Vertical Profile, Area, Eccentricity, Perimeter, Rectangularity, Gray Level, Variance, DCT, and DWT.	1000/ 288	0.52	55%
Horizontal Profile, Vertical Profile, Area, Eccentricity, Perimeter, Rectangularity, mean and mode of image, covariance, kurtosis, Aspect Ratio, correlation coefficient, Moment.	800/ 351	0.39	75%
Max and Standard Deviation of Horizontal Profile and Vertical Profile, Area, Eccentricity, Perimeter, Rectangularity, mean and mode of image, covariance, kurtosis, Aspect Ratio, correlation coefficient.	900/ 351	0.37	79%
Max, Median, and Standard Deviation of Horizontal Profile and Vertical Profile, Area, Eccentricity, Perimeter, Rectangularity, Open area, kurtosis, Moment, Aspect Ratio, Entropy, Correlation Coefficient.	900/ 351	0.252	81.8%

A. Comparison of Classifiers

Table 2 shows the computation of accuracy for RF, SVM, k-NN and LDA classifier using individual features. From Horizontal profile, Vertical profile, DCT, and Covariance features achieve high accuracy on the cost of high dimension which increases the evaluation time and complexity. On the other hand, after combining these high dimensional features accuracy starts decreasing shown in Table 3 results 1 to 4. Features like area, eccentricity, perimeter, normalized moment have low dimension but give good accuracy as compared to other features shown in table 2. In table 3 result 5 shows the combination of area, eccentricity, and perimeter gives better results with low dimensionality.



(a)

	Bangla	Devanagari	English	Gujarati	Gurumukhi	Kannada	Malayalam	Tamil	Telugu	Urdu	True Positive Rate False Negative Rate
Bangla	24 6.8%	5 1.4%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.3%	0 0.0%	1 0.3%	77.4% 22.6%
Devanagari	2 0.6%	19 5.4%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	90.5% 9.5%
English	0 0.0%	0 0.0%	33 9.4%	2 0.6%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	94.3% 5.7%
Gujarati	0 0.0%	0 0.0%	0 0.0%	23 6.6%	0 0.0%	1 0.3%	1 0.3%	3 0.9%	0 0.0%	1 0.3%	79.3% 20.7%
Gurumukhi	0 0.0%	2 0.6%	0 0.0%	0 0.0%	29 8.3%	0 0.0%	0 0.0%	1 0.3%	0 0.0%	0 0.0%	90.6% 9.4%
Kannada	0 0.0%	1 0.3%	3 0.9%	0 0.0%	0 0.0%	23 6.6%	4 1.1%	0 0.0%	1 0.3%	0 0.0%	71.9% 28.1%
Malayalam	0 0.0%	0 0.0%	0 0.0%	2 0.6%	0 0.0%	1 0.3%	31 8.8%	1 0.3%	0 0.0%	0 0.0%	88.6% 11.4%
Tamil	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.3%	2 0.6%	31 8.8%	0 0.0%	0 0.0%	91.2% 8.8%
Telugu	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	13 3.7%	8 2.3%	2 0.6%	39 11.1%	0 0.0%	62.9% 37.1%
Urdu	1 0.3%	1 0.3%	1 0.3%	1 0.3%	0 0.0%	0 0.0%	1 0.3%	0 0.0%	0 0.0%	35 10.0%	87.5% 12.5%
Positive Prediction Rate False Discovery Rate	88.9% 11.1%	67.9% 32.1%	89.2% 10.8%	82.1% 17.9%	100% 0.0%	59.0% 41.0%	66.0% 34.0%	79.5% 20.5%	97.5% 2.5%	94.6% 5.4%	81.8% 18.2%

(b)

Fig. 5 (a) OOB error rate. (b) Confusion Matrix of RF algorithm.

Table 3 shows the experimental results of various feature selection and its corresponding accuracy obtained during this study. Result number 10 shows proper feature selection to obtain the high accuracy for all four classifiers i.e. 80.9%, 77.8%, 73.5%, and 65.5% for RF, SVM, k-NN and LDA respectively and confusion matrix for these results shown in Fig. 6. In that all classifiers, LDA gives lest accuracy due to improper training. It has a limitation of dimensionality due to covariance matrix phenomenon used for classification which affects the training of the classifier. Form that all results shown in table 2 and 3, we found that RF is much better than other classifiers in case of dimensionality and time required for evaluation. The number of scripts, the similarity between various scripts, unconstrained nature (font, style, strokes, etc.) of handwriting enhanced the complexity of the whole system which affects the efficiency of the system.

V. CONCLUSION

The present work demonstrates the effectiveness of different features and classifier algorithm (i.e. RF, SVM, k-NN, LDA) integration from computer vision for Indian handwritten script identification. To increase the performance of RF classification algorithm may use 64% of data for training and the remaining 36% for testing due to the occurrence of OOB error rate which is 0.36. Variety in writer, gender, profession, and age introduced unconstrained nature in handwriting which is make script identification challenging. Features like horizontal and vertical profile projection and covariance achieved high accuracy separately. But when integrated with other features accuracy decreases and evaluation time increases due to the high dimensionality of these features. And versatile nature like standard deviation and maximum value of both profile are better for handwritten script identification due to unconstrained nature.

Classifiers have certain limitation such as LDA is not suitable for high dimension features. RF performs better in high dimension features and nonlinear data. k-NN is unable to identify the scripts which are not used during training. SVM can perform better in low dimension and linear data. Furthermore, image details (i.e. dimension, planes, color format, etc.), classifier specifications, feature vector dimension, etc. are important factors for efficient script identification system. For line-wise segmentation, vertical projection profile is not advantageous. But its standard deviation and max value are most preferable for script identification.

Table 2. Individual Evaluation of Features for Clarifiers

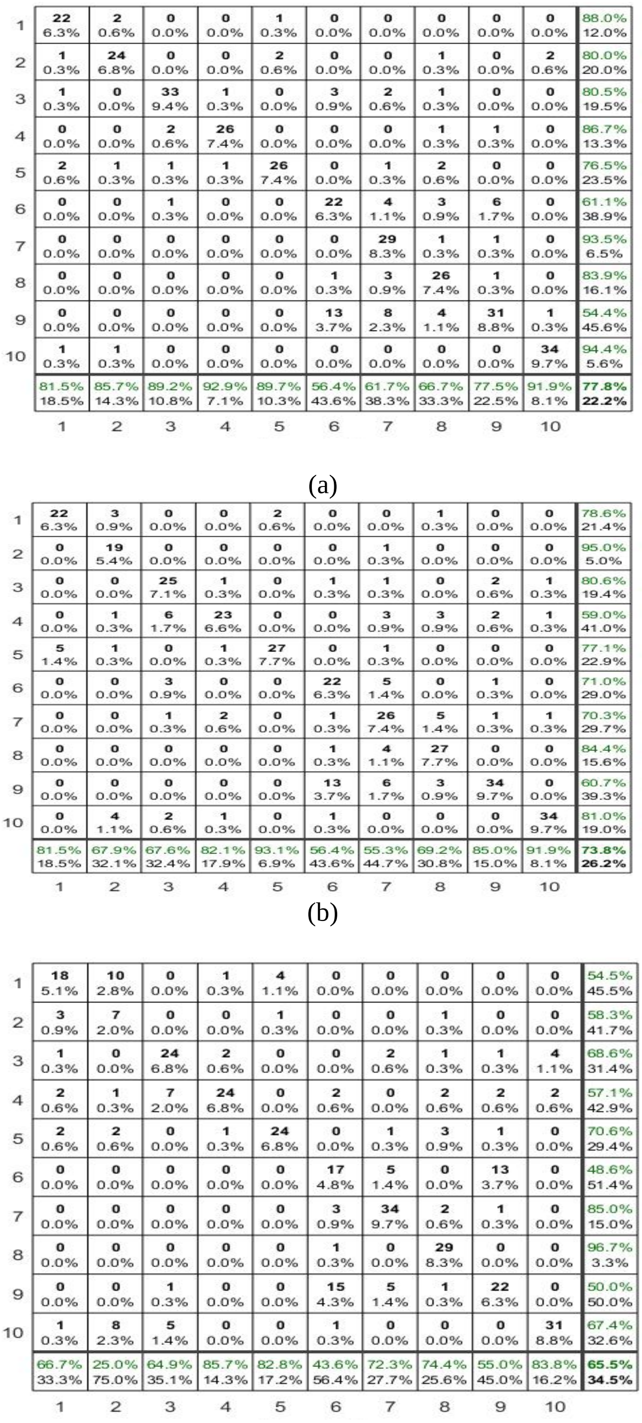
Features	RF	SVM	k-NN	LDA
Moment	37.9	31.1	38.7	24.2
Horizontal Profile	49	46.7	43.6	19.9
Standard deviation H	21.7	21.1	20.5	19.4
Vertical Profile	52.1	47	41	29.9
Standard deviation V	23.6	22.8	24.5	23.6
Rectangularity	17.4	16.5	18.2	12.8
Circularity	25.4	20.2	25.9	18.5
Aspect Ratio	22.8	29.3	24.5	26.2
Standard Deviation	21.7	20.8	20.8	17.1
Variance	21.7	20.8	20.8	18.2
Covariance	50.4	42.2	41	26.2
Correlation	14.8	15.4	14.8	12.3
Area	50.7	40.7	50.1	33
Eccentricity	38.2	35	42.5	29.6
Perimeter	55.3	47.3	54.1	41.9
Kurtosis	24.5	21.1	15.4	20.2
Entropy	21.9	20.8	20.8	17.1
Euler Number	16	14.5	11.4	14
DCT	49.3	37	29.9	36.8

Finally, the majority voting technique adopts in RF classifier enhanced the performance of the handwritten script identification system.

It obtained 81.8% accuracy and the OOB error rate of 0.252 with 73.11% of the training data by Random Forest. And 77.8%, 73.5%, and 65.5% accuracy is achieved in SVM, k-NN and LDA classifiers respectively. Finally, for better comparison and to improve the accuracy of the script identification system required the standard database. So we have to look towards this concern of creating a standard database for further study and research.

Table 3. Comparison between classifiers

Features	**RF**	**SVM**	**k-NN**	**LDA**
Horizontal Profile, Vertical Profile, Covariance	54.1	45.3	41.9	29.1
Horizontal Profile, Vertical Profile, DCT	54.4	37.3	39	32.5
Horizontal Profile, Vertical Profile	55.8	47.3	43.3	36.9
Covariance, DCT	47	36.8	33.6	27.6
Area, Eccentricity, Perimeter,	69.8	63.2	64.1	45.9
Area, Eccentricity, Perimeter, Rectangularity, Circularity, kurtosis, Moment, Aspect Ratio, Entropy, Correlation Coefficient, Moment, Gray level, Mean of Image	77.5	72.6	73.5	59.5
Rectangularity, Circularity, kurtosis, Moment, Aspect Ratio, Entropy, Correlation Coefficient, Euler Number	72.9	63	70.7	54.1
Area, Eccentricity, Perimeter, Rectangularity, Circularity, kurtosis, Moment, Aspect Ratio, Entropy, Correlation Coefficient.	78.3	71.4	72.6	56.4
Max, Median, and Standard Deviation of Horizontal Profile and Vertical Profile, Area, Eccentricity, Perimeter, Rectangularity, Circularity, Open area, kurtosis, Moment, Aspect Ratio, Entropy, Euler Number, Correlation Coefficient.	80.9	76.1	72.9	65.5
Max, Median, and Standard Deviation of Horizontal Profile and Vertical Profile, Area, Eccentricity, Perimeter, Rectangularity, Circularity, kurtosis, Normalized Moment, Aspect Ratio, Euler Number, Correlation Coefficient, Gray Level.	**81.8**	**77.8**	**73.8**	**64.1**



1	22 6.3%	2 0.6%	0 0.0%	0 0.0%	1 0.3%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	88.0% 12.0%
2	1 0.3%	24 6.8%	0 0.0%	0 0.0%	2 0.6%	0 0.0%	0 0.0%	1 0.3%	0 0.0%	2 0.6%	80.0% 20.0%
3	1 0.3%	0 0.0%	33 9.4%	1 0.3%	0 0.0%	3 0.9%	2 0.6%	1 0.3%	0 0.0%	0 0.0%	80.5% 19.5%
4	0 0.0%	0 0.0%	2 0.6%	26 7.4%	0 0.0%	0 0.0%	0 0.0%	1 0.3%	1 0.3%	0 0.0%	86.7% 13.3%
5	2 0.6%	1 0.3%	1 0.3%	1 0.3%	26 7.4%	0 0.0%	1 0.3%	2 0.6%	0 0.0%	0 0.0%	76.5% 23.5%
6	0 0.0%	0 0.0%	1 0.3%	0 0.0%	0 0.0%	22 6.3%	4 1.1%	3 0.9%	6 1.7%	0 0.0%	61.1% 38.9%
7	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	29 8.3%	1 0.3%	1 0.3%	0 0.0%	93.5% 6.5%
8	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.3%	3 0.9%	26 7.4%	1 0.3%	0 0.0%	83.9% 16.1%
9	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	13 3.7%	8 2.3%	4 1.1%	31 8.8%	1 0.3%	54.4% 45.6%
10	1 0.3%	1 0.3%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	34 9.7%	94.4% 5.6%
	81.5% 18.5%	85.7% 14.3%	89.2% 10.8%	92.9% 7.1%	89.7% 10.3%	56.4% 43.6%	61.7% 38.3%	66.7% 33.3%	77.5% 22.5%	91.9% 8.1%	**77.8%** **22.2%**
	1	2	3	4	5	6	7	8	9	10	

(a)

1	22 6.3%	3 0.9%	0 0.0%	0 0.0%	2 0.6%	0 0.0%	0 0.0%	1 0.3%	0 0.0%	0 0.0%	78.6% 21.4%
2	0 0.0%	19 5.4%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.3%	0 0.0%	0 0.0%	0 0.0%	95.0% 5.0%
3	0 0.0%	0 0.0%	25 7.1%	1 0.3%	0 0.0%	1 0.3%	1 0.3%	0 0.0%	2 0.6%	1 0.3%	80.6% 19.4%
4	0 0.0%	1 0.3%	6 1.7%	23 6.6%	0 0.0%	0 0.0%	3 0.9%	3 0.9%	2 0.6%	1 0.3%	59.0% 41.0%
5	5 1.4%	1 0.3%	0 0.0%	1 0.3%	27 7.7%	0 0.0%	1 0.3%	0 0.0%	0 0.0%	0 0.0%	77.1% 22.9%
6	0 0.0%	0 0.0%	3 0.9%	0 0.0%	0 0.0%	22 6.3%	5 1.4%	0 0.0%	1 0.3%	0 0.0%	71.0% 29.0%
7	0 0.0%	0 0.0%	1 0.3%	2 0.6%	0 0.0%	1 0.3%	26 7.4%	5 1.4%	1 0.3%	1 0.3%	70.3% 29.7%
8	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.3%	4 1.1%	27 7.7%	0 0.0%	0 0.0%	84.4% 15.6%
9	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	13 3.7%	6 1.7%	3 0.9%	34 9.7%	0 0.0%	60.7% 39.3%
10	0 0.0%	4 1.1%	2 0.6%	1 0.3%	0 0.0%	1 0.3%	0 0.0%	0 0.0%	0 0.0%	34 9.7%	81.0% 19.0%
	81.5% 18.5%	67.9% 32.1%	67.6% 32.4%	82.1% 17.9%	93.1% 6.9%	56.4% 43.6%	55.3% 44.7%	69.2% 30.8%	85.0% 15.0%	91.9% 8.1%	**73.8%** **26.2%**
	1	2	3	4	5	6	7	8	9	10	

(b)

1	18 5.1%	10 2.8%	0 0.0%	1 0.3%	4 1.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	54.5% 45.5%
2	3 0.9%	7 2.0%	0 0.0%	0 0.0%	1 0.3%	0 0.0%	0 0.0%	1 0.3%	0 0.0%	0 0.0%	58.3% 41.7%
3	1 0.3%	0 0.0%	24 6.8%	2 0.6%	0 0.0%	0 0.0%	2 0.6%	1 0.3%	1 0.3%	4 1.1%	68.6% 31.4%
4	2 0.6%	1 0.3%	7 2.0%	24 6.8%	0 0.0%	2 0.6%	0 0.0%	2 0.6%	2 0.6%	2 0.6%	57.1% 42.9%
5	2 0.6%	2 0.6%	0 0.0%	1 0.3%	24 6.8%	0 0.0%	1 0.3%	3 0.9%	1 0.3%	0 0.0%	70.6% 29.4%
6	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	17 4.8%	5 1.4%	0 0.0%	13 3.7%	0 0.0%	48.6% 51.4%
7	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	3 0.9%	34 9.7%	2 0.6%	1 0.3%	0 0.0%	85.0% 15.0%
8	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.3%	0 0.0%	29 8.3%	0 0.0%	0 0.0%	96.7% 3.3%
9	0 0.0%	0 0.0%	1 0.3%	0 0.0%	0 0.0%	15 4.3%	5 1.4%	1 0.3%	22 6.3%	0 0.0%	50.0% 50.0%
10	1 0.3%	8 2.3%	5 1.4%	0 0.0%	0 0.0%	1 0.3%	0 0.0%	0 0.0%	0 0.0%	31 8.8%	67.4% 32.6%
	66.7% 33.3%	25.0% 75.0%	64.9% 35.1%	85.7% 14.3%	82.8% 17.2%	43.6% 56.4%	72.3% 27.7%	74.4% 25.6%	55.0% 45.0%	83.8% 16.2%	**65.5%** **34.5%**
	1	2	3	4	5	6	7	8	9	10	

(c)

Fig. 6 Confusion matrix of Classifier (a) SVM (b) k_NN (c) LDA

ACKNOWLEDGMENT

I am truly grateful to my supervisor, **Dr. D. V. Rojatkar**, who gave me a chance to work on the **Indian handwritten script identification** project. His expertise on machine learning, which he constantly shared with me, effectively contributed to the success of my PG project. From the first until the last day of my PG time, he was very friendly and always had a positive motivation. I have learned a lot of things from him. His advice and supervision literally helped me growing up more in the academic pathway.

I also want to say special thanks to my PG fellow, UG students and Faculty member of the Department of Electronics Engineering, who gave me some useful advice and ideas about my Indian handwritten script identification project. I am thankful to everyone in my Electronics Department for believing in me and for keeping pushing me forward.

Finally but most importantly, I would like to express my gratitude to my family, especially my mother, **Chanda**, my father, **Prakashrao** and my younger brother **Pawan**, who always gave me unconditional love and support throughout my study.

REFERENCES

1. Dinesh V. Rojatkat, Krushna D. Chinchkhede, G.G. Sarate, “Design and Analysis of LRTB feature based Classifier applied to Handwritten Devnagari Characters:A Neural Network Approach,” *International Conference on Advances in Computing, Communications and Informatics (ICACCI)* , 2013.
2. Dinesh V Rojatkar, Krushna D Chinchkhede, GG Sarate, “Handwritten Devnagari consonants recognition using MLPNN with five fold cross validation,” *International Conference on Circuits, Power and Computing Technologies (ICCPCT)*, 2013.
3. Judith Hochberg, Kevin Bowers, Michael Cannon, Patrick Kelly, “ Script and language identification for handwritten document image,”. *International Journal on Document Analysis and Recognition Springer-Verlag*, 1999.
4. V. Singhal, D. Ghosh, and N. Navin, “Script Based Classification of Handwritten Text Documents in a Multi-lingual Environment,” *Proc. Intel Workshop Research Issues in Data Engineering Multi-Lingual Information Management*, 2003.
5. K.-Roy, A. Baner and U. Pal, “ A System for Word-wise Handwritten Script Identification for Indian Postal Automation,” *IEEE India Annual Conference (INDICON)*, 2004.
6. Anoop M. Namboodiri and A. K. Jain, “ Online Handwritten Script Recognition,” *IEEE Transactions*, 2004.
7. G. G. Rajput, Anita H. B., “ Handwritten Script Recognition using DCT and Wavelet Features at Block Level,” *International Journal of Computer Applications (IJCA) Special Issue on RTIPPR*, 2010.
8. Mallikarjun Hangarge and B.V. Dhandra, “Offline Handwritten Script Identification in Document Images,” *International Journal of Computer Applications* (0975 – 8887) Volume 4 – No.6, 2010.
9. Sk Md Obaidullah, Supratik Kundu Das, Kaushik Roy, “ A System for Handwritten Script Identification from Indian Document,” *Journal of Pattern Recognition Research*, 2013.
10. Rajmohan Pardeshi, B. B. Chaudhuri, Mallikarjun Hangarge, K.C. Santosh, “Automatic Handwritten Indian Scripts Identification,” *14 International Conference on Frontiers in Handwriting Recognition*, 2014.
11. G.G. Rajput, Suryakant Baburao Ummapure, “Script Identification from Handwritten Documents using SIFT.,” *IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, 2017.
12. S. Chanda and U. Pal, “ English, Devnagari and Urdu Text Identification,” *Proceedings of the International Conference on Cognition and Recognition* (2018).
13. S. Jeyalaksshmi, S. Prasanna, “Measuring distinct regions of grayscale image using pixel values,” *International Journal of Engineering & Technology*, 7 (1.1) (2018) 121-124.
14. Jamileh YousefiUniversity of Guelph, “Image Binarization using Otsu Thresholding Algorithm,” 10.13140/RG.2.1.4758.9284, April 18, 2011.
15. Priya M.S, Dr. G.M. Kadhar Nawaz, “Multilevel Image Thresholding using OTSU’s Algorithm in Image Segmentation,” *International Journal of Scientific & Engineering Research* Volume 8, Issue 5, May-2017.
16. Vijay Kumar, Priyanka Gupta, “ Importance of Statistical Measures in Digital Image Processing,” *International Journal of Emerging Technology and Advanced Engineering*, August 2012.
17. Leo Breiman, “RANDOM FORESTS,” *UC Berkeley Statistic* , January 2001.

AUTHORS PROFILE

Lalit P. Ganorkar received his B.E. and M.Tech degrees from the University of Nagpur and Amravati, in 2013 and 2019, respectively. His research interests include Artificial Intelligence, Pattern Recognition, and Machine learning. The Author published his research paper in IEEE and Springer conferences related to Indian handwritten script identification.

Dinesh V. Rojatkar received his M.E. degree in Electronics Engineering from the University of SGBAU, Amravati, India, and his Ph.D. degree in Engineering and Technology from the University of SGBAU, Amravati, India, in 2003 and 2013, respectively. From 1999 to 2018 he was at the University of Amravati and Nagpur, as Assistant Professor. He is currently an Associate Professor in the Department of Electronics Engineering, the University of SGBAU, and Amravati, India.